# A Definitions of models and algorithms we used

## A.1 Steiner tree problem

**Definition A.1** (Steiner tree)**.** *Let $G = \{V, E\}$ be an edge-weighted graph. Let $W$ be a subset of nodes in $V$. The* Steiner tree problem *is to find a tree that spans all the nodes in $W$ and its edge weight sum is minimized.*

## A.2 Small world

We now review Kleinberg's small world model (Kleinberg, 2000). We remark that Kleinberg's model allows the network to be directed. Here, we shall use a straightforward way to convert a directed graph into an undirected one: there is an undirected edge between $u$ and $v$ if and only if $(u, v)$ or $(v, u)$ is in the directed grpah. Thus, Kleinberg's small world model can be described as follows.

The set of nodes reside in a two-dimensional lattice points $\{(i, j) : i, j \in \{1, ..., \sqrt{n}\}\}$. The *lattice distance* is defined as $d((i, j), (k, \ell)) = |k - i| + |\ell - j|$. Let $r = 2$ and $q = \Theta(\log n)$ be a normalization term. There are two types of edges in the small world graph:

- *local edges:* if $d(u, v) \leq 1$, then there is an edge between $d(u, v)$.
- *long range edges:* if $d(u, v) > 1$, then with probability $q^{-1} \cdot d^{-r}(u, v)$ there is an edge between $u$ and $v$, which is independent to the rest of the edges.

We shall define $d \triangleq \max_v \mathrm{E}[\text{degree}(v)]$. Since $q = \Theta(\log n)$, $d$ is also a constant.

## A.3 Chung and Lu's random graph

Next, we describe Chung and Lu's random graph model (Chung and Lu, 2002). In this model, we are given an expected degree sequence $\mathbf{w} = (w_1, ..., w_n)$, in which $w_i$ represents the expected degree for $v_i$. The probability that there is an edge between $v_i$ and $v_j$ is $w_i w_j \rho$ for $\rho = \frac{1}{\sum_i w_i}$, which is independent to other edges. Furthermore, we shall assume the distribution of $\{w_1, ..., w_n\}$ follows a power law distribution with exponent parameter between 2 and 3.

# B Proof of Theorem 3.1

We shall partition the support of the signals into three regions: the *high interval* $H = [1, 2 - \gamma]$, the *median interval* $M = (1 - \gamma, 1)$, and the *low interval* $L = [0, 1 - \gamma]$. When a node's signal is in $H$, the node is in $S$; when a node's signal is in $L$, the node is not in $S$. Thus, a detector's only goal is to identify the nodes from $S$ among those nodes with their signals in $M$.

On the other hand, observe that conditioned on a node's signal is in $M$, the distribution of the signal is uniform in $(1 - \gamma, 1)$. This always holds regardless whether the node is in $S$ or not. Therefore, no algorithm can do better than random guess in the region $M$. It remains to analyze the performance of an algorithm that only makes random guesses. One can see that with high probability: 1. the number of signals in $M$ is $\Theta(\gamma n)$. 2. the number of nodes in $S$ with signals in $M$ is $\Theta(\gamma k)$. Thus, on average, a node in $S$ will be picked up with probability $\Theta\left(\frac{\gamma k}{\gamma n}\right) = \Theta(k/n)$. Since we can select up to $O((\alpha - 1)k)$ nodes in the interval $M$ (with high probability), the total number of nodes from $S$ that will finally be in $\hat{S}$ is $O(k^2/n)$. Thus,

$$\mathrm{E}\left[|S - S'|\right] = (1 - o(1))\gamma k - O\left(\frac{k^2}{n}\right) = (1 - o(1))\gamma k$$

for sufficiently large $n$ and $k$.

# C Warmup: the algorithm for a line case.

We now give a solution for the case where the network forms a line, *i.e.,* $E = \{\{v_i, v_{i+1}\} : i < n\}$. While a line graph is not a realistic model for social or biological networks, analyzing this simple example helps us to understand how we may improve the performance of a detector algorithm by utilizing the information of the network structure.

**Lemma C.1.** *Let $\gamma$ be an arbitrary constant. Consider the community detection problem, in which the underlying graph is a line and $k = o(n)$ is polynomial in $n$. There exists an efficient algorithm that returns a set $\hat{S}$ of size $k$ and $|S - \hat{S}| = o(k)$ whp. In other words, for any $\gamma$, there exists a $(1, \gamma)$-detector.*

*Proof.* Observe that the subgraph induced by $S$ has to be connected. Hence, there exists a $j$ such that $S = \{v_j, v_{j+1}, ..., v_{j+k-1}\}$. Our algorithm works as follows: enumerate through all possible connected subgraphs of size $k$. Output an arbitrary one such that it contains at least $(1-\gamma)k-(\log n)\sqrt{k}$ signals in the high interval $H$. Also in what follows, we say a connected subgraph of size $k$ *passes the threshold test* if and only if it contains at least $(1-\gamma)k-(\log n)\sqrt{k}$ signals in the high interval.

We need to show the following two events hold with high probability

- *Event 1:* The subgraph induced by $S$ passes the threshold test.
- *Event 2:* There exists no connected subgraph $S'$ such that $|S'| = k$, $|S - S'| = \Omega(k)$, and $S'$ passes the threshold tests.

Event 1 implies that with high probability our algorithm will return at least one subgraph. Event 2 implies that with high probability the node set returned by our algorithm will be almost the same as $S$.

We may apply a standard Chernoff bound to prove that event 1 happens with high probability. Specifically,

$$\Pr[S \text{ fails the threshold test}] \leq \exp\left(-\frac{\log^2 n}{3}\right) \leq \frac{1}{n^2} \tag{4}$$

for sufficiently large $n$.

We next move to analyze the second event. Let $\epsilon$ be an arbitrarily small constant. We shall first show the probability that there exists a specific connected subset $S'$ such that $|S - S'| \geq \epsilon k$ is small. Then we shall use a union bound to argue whp no bad $S'$ will pass the threshold test.

Let $S'$ be an arbitrary connected subset of size $k$ such that $|S - S'| \geq \epsilon k$. In expectation, the total number of nodes in $S$ that have signals in $H$ is $(1-\epsilon)(1-\gamma)k$. By using a Chernoff bound again, the probability that $S'$ will have more than $(1-\gamma)k - (\log n)\sqrt{k}$ nodes in $H$ is $\leq \frac{1}{n^3}$ (no effort was made to optimize this bound). Finally, since there are at most $(n-k+1)$ connected subgraphs of size $k$, we may apply a union bound and get that the probability event 2 happens is $\leq \frac{n-k+1}{n^3} \leq 1/n^2$.

To sum up, with probability $1 - \frac{2}{n^2}$, we will find a set $\hat{S}$ of size $k$ such that $|S - \hat{S}| = o(k)$. □

## D Proof of Lemma 4.4

We need to prove two directions: 1. a **MinConnect** problem can be reduced to a Steiner tree problem with uniform edge weights. 2. a Steiner tree problem with uniform edge weights can be reduced to a **MinConnect** problem.

We use the following observation for the analysis for both directions: in an arbitrary graph with uniform edge weights, the cost of any spanning tree for a connected subset of size $k$ is $k-1$. This observation is true because of the simple fact that a tree of size $\ell$ contains $\ell - 1$ edges for arbitrary $\ell$.

Now both directions are straightforward. Let $W$ be the set of nodes that need to be included in the **MinConnect** problem. To reduce the **MinConnect** problem to a Steiner tree problem, we require the Steiner tree to cover all the nodes in $W$ in the same graph. Because of the above observation, a Steiner tree that covers $W$ with minimum number of edges is also a minimum connected subgraph that covers $W$.

We can prove the other direction in a similar fashion.

## E Proof of Proposition 4.6

We shall show that with probability $\left(\frac{\tau\ell}{n}\right)^{\ell-1}$ a random subset $R$ of size $\ell$ is connected, where $\tau$ is a suitable constant (the probability is over both the generative model and the choice of the random subset). Then we can show that the expected number of subgraphs of size $\ell$ is $\binom{n}{\ell}\left(\frac{\tau\ell}{n}\right)^{\ell-1} = n\tau_0^\ell$ for another constant $\tau_0$. We may next apply a Markov inequality to get the Proposition.

First, we shall imagine the set $R$ is sampled in a sequential manner, *i.e.,* the first node $v_1$ from $R$ is sampled uniformly from $V$. Then the second node $v_2$ is sampled uniformly from $V/\{v_1\}$. And in general the $i$-th node $v_i$ in $R$ is sampled uniformly from $V/\{v_1, ..., v_{i-1}\}$. After $\ell$ samplings, we would get a set $R$ of size $\ell$ that is sampled uniformly among all the nodes. The reason we introduce this scheme is to couple the sampling of $R$ with the generative model itself, by using the principle of deferred decision over long range edges: the existence of a long range edge is not revealed until we need it to make conclusions on connectivity. In other words, we want to test the connectivity of $R$ sequentially when each member node of $R$ and its neighbors are revealed to us step-by-step. The bottom line is that we need not know all the member nodes of $R$ to conclude on connectivity.

To carry out our scheme, we introduce a coupled branching process that is represented by a stochastic tree $\{T(t)\}_{t\geq 1}$. This tree traces the current revealed nodes in $V(R)$. It is rooted at $v_1$ and grows over time according to the following procedure. First, let us introduce a labeling notion to help define our growing process: at each time $t$, we call any node in $V$ *dead* if its long-range neighbors are revealed to us, and *active* otherwise. The notion of dead and active can apply to any nodes in $V$.

We start from a uniformly sampled node $v_1$ in $V$, decide $v_1 \in R$, and reveal all its long-range neighbors (in addition to the four local neighbors) via the generative model. Then $T(1)$ contains only $v_1$ and $v_1$ is labelled dead. To grow from $T(t)$ to $T(t+1)$, we implement the following steps:

- *Step 1.* Pick an arbitrary active node $v$ in $V(T(t))$ and reveal all its neighbors. The active node is then labeled as dead.
- *Step 2.* Decide the set of nodes from $R/V(T(t))$ that are adjacent to $v$ by using this observation: the distribution of $R/(V(T(t)))$ is uniform over $V/\left(V(T(t)) \cup \text{Dead}(t)\right)$, where $\text{Dead}(t)$ is the

union of all neighbors of all the dead nodes at time $t$.

- *Step 3.* Expand $T(t)$ to $T(t+1)$ by adding all the new neighbors of $v$ that are in $R$.

Note that $V(T(t)) \in R$ by construction. Moreover, whenever the tree $T(t)$ stops growing before $|T(t)|$ reaches $\ell$, we conclude that the current sample of $R$ is not connected.

We next show by induction over $\ell$ that $\Pr[|T(t)| = k] \leq \left(\frac{\tau\ell}{n}\right)^{\ell-1}$. One potential obstacle in our analysis is that in the second step of the above procedure, the set $\mathrm{Dead}(t) \cup V(T(t))$ evolves in a complex manner, and it is not straightforward to find the neighbors of $v$ that are in $R$ over time. On the other hand, it is not difficult to show that $\Pr[|\mathrm{Dead}(t)| \geq \frac{n^{2/3}}{2}] \leq 2^{-\frac{n^{2/3}}{2}}$ at any moment (by using a Chernoff bound). In the rest of our analysis, we shall silently assume that $\mathrm{Dead}(t) \geq \frac{n^{2/3}}{2}$ does not happen. This assumption will result in an additive error of order $2^{-\frac{n^{2/3}}{2}}$ for the probability quantities we are interested in, which is an asymptotically small term. Since our goal is to give an upper bound on the probability that $|T(t)|$ eventually reaches $\ell$, we can imagine the branching process works as follows: there is an (imaginary) adversary that decides how $T(t)$ grows and tries to maximize the probability that $|T(t)|$ eventually reaches $\ell$. The adversary basically has to follow the above three-step procedure to grow $T(t)$, but in the second step the adversary needs to decide which set of nodes that is not allowed to choose (instead of using $\mathrm{Dead}(t) \cup V(T(t))$) as long as $|\mathrm{Dead}(t)|$ is sufficiently small, *i.e.,* $|\mathrm{Dead}(t)| \leq \frac{n^{2/3}}{2}$. When the adversary maximizes the probability that $|T(t)|$ eventually reaches $\ell$, such quantity is also an upper bound on the same probability for the original process.

Thus, our goal is to understand the probability that $|T(t)|$ hits $\ell$ under the adversarial setting. Specifically, we want to prove the following statement.

*Let $\ell \leq n^{1/3}$ and let $G$ be a small world graph of size $n$. Let $F$ be an arbitrary set of size $\frac{n}{2} - \frac{\ell}{2} \cdot n^{2/3}$. We shall refer $F$ as the* forbidden set*. Let $R$ be a random subgraph from $V/F$ of size $\ell$. Let $P(\ell, n)$ be the probability that the coupled branching process $T(t)$ reaches $\ell$ nodes eventually. Then*

$$P(\ell, n) \leq \left(\frac{\tau\ell}{n}\right)^{\ell-1}. \tag{5}$$

Roughly speaking, the forbidden set $F$ allows the adversary to choose the set $\mathrm{Dead}(t) \cup V(T(t))$ over time. Also, notice that the set $F$ has to shrink as $\ell$ grows. Imposing such a technical assumption will makes the recursive analysis easier at the cost of weakening bound on $k$. Also, we shall refer to the branching process in which the adversary controls the set $F$ as $T'(t)$.

The base case where $\ell = 1, 2$ is straightforward. Let us now move to the induction step for computing $P(\ell+1, n)$ when $P(1, n), ..., P(\ell, n)$ satisfy (5). Let $v_1$ be the first node in $R$. Let us define the random variables $X$ and $Y$ as follows: $X$ is the total number of neighbors of $v_1$ and $Y$ is the total number of nodes from $R$ that are adjacent to $v_1$. Notice that $\mathrm{E}[X_1] \leq 4 + d$ (4 is the number of direct neighbors and $d$ is the maximum expected number of long range edges among all the nodes). By using a Chernoff bound, we also have $\Pr[X_1 \geq g] \leq 2^{-g}$, when $g$ is a sufficiently large constant (Chernoff bound is applicable because the number of long range edges can be expressed as the sum of independent variables).

Furthermore, we may also compute $\Pr[Y = j]$ asymptotically. Specifically, we have the following lemma.

**Lemma E.1.** *Let $Y$ be the variable defined above. For any $i \leq \ell$, there exists a constant $c_2$ such that*

$$\Pr[Y = i] \leq \left(\frac{c_2\ell}{n}\right)^{i}. \tag{6}$$

*Proof of Lemma E.1.* Recall that we let $X$ be the number of nodes that are adjacent to $v_1$. Also, recall that $\Pr[X > g] < 2^{-g}$, when $g$ is a sufficiently large constant. We can also write $\Pr[X > g] \leq c_4 \cdot 2^{-g}$ for a sufficiently large $c_4$.

We have

$$\begin{aligned} \Pr[Y = i] &= \textstyle\sum_{1 \leq j \leq n} \Pr[Y = i | X = j] \Pr[X = j] \\ &\leq c_4 \textstyle\sum_{i \leq j \leq n} \frac{\binom{j}{i}\binom{n_0 - j}{\ell - i}}{\binom{n_0}{\ell}} \cdot 2^{-j}, \end{aligned} \tag{7}$$

where $n_0$ is $n - |F \cup \{v_1\}|$ is the total number of nodes that $R$ may choose from. Notice that $n_0 = \Theta(n)$ .

Let us focus on the terms $\frac{\binom{j}{i}\binom{n_0-j}{\ell-i}}{\binom{n_0}{\ell}} \cdot 2^{-j}$. One can see that when $j \geq 3i$, the terms $\frac{\binom{j}{i}\binom{n_0-j}{\ell-i}}{\binom{n_0}{\ell}} \cdot 2^{-j}$ decrease more sharply than a geometric progression with ratio $3/4$. Thus, the sum of the first $3i$ terms is the dominating term. We now give an upper bound on $\frac{\binom{j}{i}\binom{n_0-j}{\ell-i}}{\binom{n_0}{\ell}} \cdot 2^{-j}$ for the case $j \leq 3i$. Let $j = \alpha i$, where

$\alpha \le 3$, we have

$$\begin{aligned}
&\frac{\binom{\alpha i}{i}\binom{n_0-\alpha i}{\ell-i}}{\binom{n_0}{\ell}} \\
\le\ & \frac{c_5}{\sqrt{\ell}} \frac{\alpha^i \frac{(n_0-\alpha i)^{n_0-\alpha i}}{(\ell-i)^{\ell-i}(n_0-\ell-(\alpha-1)i)^{n_0-\ell-(\alpha-1)i}}}{\frac{n_0^{n_0}}{\ell^\ell (n_0-\ell)^{n_0-\ell}}} \\
&\qquad (c_5 \text{ is a sufficiently large constant.}) \\
\le\ & c_5 \cdot \alpha^i \frac{\{(n_0-\alpha i)(n_0-\ell)\}^{n_0-\ell-(1-\alpha)i}}{\{(n_0-\ell-(\alpha-1)i)n_0\}^{n_0-\ell-(\alpha-1)i}} \\
&\cdot \frac{\{\ell(n_0-\alpha i)\}^{\ell-i}}{\{(\ell-i)n_0\}^{\ell-i}} \cdot \frac{(n_0-\ell)^{(\alpha-1)i}}{n_0^{(\alpha-1)i}} \frac{\ell^i}{n_0^i}.
\end{aligned}$$

We can see that $(n_0-\alpha i)(n_0-\ell) \le (n_0-\ell-(\alpha-1)i)n$, $\ell(n_0-\alpha i) \le (\ell-i)n_0$, and $n_0-\ell \le n_0$. Thus,

$$\frac{\binom{\alpha i}{i}\binom{n_0-\alpha i}{\ell-i}}{\binom{n_0}{\ell}} \le \left(\frac{c_6 \ell}{n}\right)^2.$$

Together with (7), we finish the proof of Lemma E.1 □

Next, by the law of total probability, we also have $\Pr[|T'(t)| = \ell+1]$ equals to $\sum_{1\le i\le n} \Pr[|T'(t)| = \ell+1 | Y=i] \Pr[Y=i]$. We first walk through the analysis for $\Pr[|T'(t)| = \ell+1 | Y=1]$ and $\Pr[|T'(t)| = \ell+1 | Y=2]$. Then we give an asymptotic bound on $\Pr[|T'(t)| = \ell+1 | Y=i]$ for general $i$.

Let us start with the case $Y=1$. Let $v_2$ be the node in $S$ that is connected with $v_1$. Here, $\Pr[|T'(t)| = \ell+1 \mid Y=1]$ reduces to a case that is covered by the induction hypothesis, *i.e.,* the tree $T'(t)$ contains $\ell+1$ nodes if and only if the subtree rooted at $v_2$ contains $\ell$ nodes. This $v_2$-rooted subtree is another branching process coupled with $\ell$ nodes, which are uniformly chosen from $V/(\{v_1\} \cup \Gamma(v_1) \cup F)$, where $\Gamma(v_1)$ is the set of $v_1$'s neigbhors. Thus, by induction hyptothesis, we have $\Pr[|T'(t)| = \ell+1 \mid Y=1] \le P(\ell,n) \le \left(\frac{\tau\ell}{n}\right)^{\ell-1}$.

Next, let us move to the case for $Y=2$. Let $v_2$ and $v_3$ be the nodes in $S$ that are connected with $v_1$. We may first grow the tree $T'(t)$ from $v_2$, then grow the tree from $v_3$. At the end, let $i$ be the number of children of $v_2$ (and $v_2$ itself). The number of children of $v_3$ (and $v_3$ itself) is thus $\ell-i$. One may check that both subprocesses can be understood by using induction hypothesis (we also need to check that the size of the forbidden set does not violate the requirements in the induction hypothesis; but this is straightforward because we assumed that the total number of neighbors of $R$ is less than $n^{2/3}/2$).

Thus, we have

$$\begin{aligned}
&\Pr[|T(t)| = \ell+1 | Y=2] \\
=\ & \sum_{i=1-1}^{\ell} \Big( \Pr[v_2\text{-rooted tree connected}, \\
&\qquad v_3\text{-rooted tree connected} \mid L=i, Y=2] \\
&\qquad \Pr[L=i \mid Y=2] \Big) \\
\le\ & \sum_{i=1}^{\ell-1} P(i,n) P(\ell-i,n) \binom{\ell-2}{i-1} \\
\le\ & \sum_{i=1}^{\ell-1} \left(\frac{\tau i}{n}\right)^{i-1} \left(\frac{\tau(\ell-i)}{n}\right)^{\ell-i-1} \binom{\ell-2}{i-1} \quad \text{(Induction)} \\
=\ & \frac{\tau^{\ell-2}}{n^{\ell-2}} \sum_{i=1}^{\ell-1} i^{i-1} (\ell-i)^{\ell-i-1} \binom{\ell-2}{i-1}
\end{aligned}$$

Let us define $a_j = j^{j-1}(\ell-j)^{\ell-j-1}$. Notice that $a_j = a_{\ell-j}$. So we have

$$\sum_{1\le j\le \ell} a_j \le 2 \sum_{1\le j\le \ell/2} a_j.$$

We may continue to compute the sum of $a_j$'s:

$$\begin{aligned}
&\sum_{1\le j\le \ell/2} a_j \\
=\ & \sum_{1\le j\le \ell/2} j^{j-1}(\ell-j)^{\ell-j-1} \binom{\ell-2}{j-1} \\
\le\ & \sum_{1\le j\le \ell/2} j^{j-1} \ell^{\ell-j-1} \left(\frac{\ell-j-1}{\ell}\right)^{\ell-j-1} \binom{\ell-2}{j-1} \\
\le\ & e \sum_{1\le j\le \ell/2} \ell^{j-1} \ell^{\ell-j-1} \left(\frac{\ell-j-1}{\ell}\right)^{\ell-j-1} \\
\le\ & c_3 \ell^{\ell-2}
\end{aligned}$$

for some constant $c_3$. The second inequality uses the fact that

$$\begin{aligned}
j^{j-1}\binom{\ell-2}{j-1} &\le j^{j-1} \left(\frac{\ell-2}{j-1}\right)^{j-1} \\
&= j^{j-1} \left(\frac{\ell-2}{j}\right)^{j-1} \left(\frac{j}{j-1}\right)^{j} \\
&\le e\ell^{j-1}
\end{aligned}$$

Thus, we have

$$\Pr[|T(t)| = \ell+1 \mid Y=2] = c_3 \left(\frac{\tau\ell}{n}\right)^{\ell-2}. \quad (8)$$

In general, we may define the $\ell$-hitting probability $P_j(\ell,n)$ for a "branching forest" that grows from $j$

different roots. When $j = 1$, it corresponds with the case for $Y = 1$, *i.e.,* $P_1(\ell, n) = P(\ell, n)$. When $j = 2$, (8) gives us $P_2(\ell, n) \leq c_3 \left(\frac{\tau\ell}{n}\right)^{\ell-2}$.

We now compute the general form of $P_j(\ell, n)$. Specifically, we shall show by induction that

$$P_j(\ell, n) \leq c_3^{j-1} \left(\frac{\tau\ell}{n}\right)^{\ell-j}. \tag{9}$$

The base cases for $j = 1, 2$ are already analyzed above. We now move to the induction case. We have the following recursive relation:

$$P_j(\ell, n) \leq \sum_{1\leq i\leq \ell-j+1} P_1(i, n) P_{j-1}(\ell - i, n) \binom{\ell - j}{i - 1}.$$

Thus, we have

$$\begin{aligned} & P_j(\ell, n) \\ \leq & \sum_{1\leq i\leq \ell-j+1} \left(\tfrac{\tau i}{n}\right)^{i-1} \left(\tfrac{\ell-i}{n}\right)^{\ell-i-j+1} c_3^{j-2} \tau^{\ell-i-j+1} \\ & \qquad \cdot \tbinom{\ell-j}{i-1} \\ = & \left(\tfrac{\tau}{n}\right)^{\ell-j} c_3^{j-2} \left(\sum_{1\leq i\leq \ell-j+1} i^{i-1} (\ell - i)^{\ell-i-j+1} \tbinom{\ell-j}{i-1}\right) \end{aligned} \tag{10}$$

It remains to analyze the term $\sum_{1\leq i\leq \ell-j+1} i^{i-1}(\ell - i)^{\ell-i-j+1}\binom{\ell-j}{i-1}$. Via some straightforward manipulation, we have

$$\sum_{1\leq i\leq \ell-j+1} i^{i-1}(\ell - i)^{\ell-i-j+1} \binom{\ell - j}{i - 1} \leq c_3 \ell^{\ell-j} \tag{11}$$

for some sufficiently large $c_3$. (11) and (10) together give us (9). Now we are ready to compute $P(\ell + 1, n)$ (and thus $\Pr[|T'(t)| = \ell + 1]$):

$$\begin{aligned} P(\ell + 1, n) &= \sum_{1\leq j\leq n} P(\ell + 1, n \mid Y = j) \Pr[Y = j] \\ &\leq \sum_{1\leq j\leq n} \left(\frac{\tau\ell}{n}\right)^{\ell-j} c_3^{j-1} \left(\frac{c_2\ell}{n}\right)^{j} \end{aligned}$$

When $\tau \geq 4c_2c_3$, we have

$$\begin{aligned} & \sum_{1\leq j\leq n} \left(\frac{\tau\ell}{n}\right)^{\ell-j} c_3^{j-1} \left(\frac{c_2\ell}{n}\right)^{j} \\ \leq & \sum_{1\leq j\leq n} \left(\frac{\tau\ell}{n}\right)^{\ell} 2^{-i} \leq \left(\frac{\tau\ell}{n}\right)^{\ell}. \end{aligned}$$

This completes the proof of Proposition 4.6.

## F Proof of Proposition 4.8

We shall first describe a way to construct $S$. Then we will argue that $\gamma$ portion of the nodes in $S$ is statistically indistinguishable with a large number of nodes from $V/S$.

Recall that $w_i$ is the average degree for the node $v_i$. Wlog, we shall let $w_1 \leq w_2 \leq ... \leq w_n$, where $w_n = c_0\sqrt{n}$ for some constant $c_0$. Since the degree distribution is a power law distribution, there exists a constant $c_1$ such that for all $i \leq c_1 n$, $w_i = \Theta(1)$. Let us randomly partition the set of nodes $\{v_1, v_2, ..., v_{c_1 n}\}$ into two subsets of equal size, namely $S_1 = \{v_{i_1}, ...., v_{i_{c_1 n/2}}\}$ and $S_2 = \{v_{i'_1}, ...., v_{i'_{c_1 n/2}}\}$. Notice that for any $v_i \in S_1 \cup S_2$, we have $\Pr[\{v_i, v_n\} \in E] = w_i w_n \rho \geq \frac{c_2}{\sqrt{n}}$ for some constant $c_2$.

Next, let us construct the set $S$: we shall first let $v_n \in S$ and the rest of the nodes in $S$ will be picked up from $S_1$. Since with probability at least $\frac{c_2}{\sqrt{n}}$ there is an edge between a node in $S_1$ and $v_n$, in expectation the number of nodes that are connected with $v_n$ is $\Omega(\sqrt{n})$. Thus, with high probability we are able to find a subset of size $k - 1$ that are all connected to $v_n$.

Now we analyze an arbitrary algorithm's performance. By using a Chernoff bound, we can see that with high probability in $S$ there are $(1 - \gamma \pm o(1))k$ nodes that have signals in $H$ and $(\gamma \pm o(1))$ nodes that have signals in $M$. Let us refer to the subset of nodes in $S$ whose signals are in $M$ as $S_M$. Recall that when the algorithm does not know the network structure, it will not be able to discover most of the nodes in $S_M$. Here, we shall show that the algorithm will behave in a similar way even that it knows the network structure.

Let us focus on the nodes in $S_2$. It is strightforward to see that with high probability the number of nodes in $S_2$ that are both connected with $v_n$ and associated with signals in $M$ is at lesat $(1 - \epsilon)\gamma\sqrt{n}$ for an arbitrary constant $\epsilon$. Let us call the set of these nodes $S'_2$. The nodes in $S'_2$ are statistically indistinguishable from the nodes in $S_M$. Furthermore, the connectivity constraint is met for nodes in both sets. Thus, no algorithm can do better than randomly guessing. In other words, if an algorithm outputs a set of size $O(k)$, then the number of nodes in $S_M$ that will be included is $o(\gamma k)$. This completes the proof of Proposition 4.8.

## G Proof of Proposition 5.1

Let $\ell = |S_{\text{opt}}|$ be the size of the output. Wlog, we shall focus on the case $\ell \geq (1 + \epsilon)k$. The case for $\ell \leq (1 + \epsilon)k$ can be analyzed in a similar manner. Recall that $\Phi(x)$ is the cdf for the Gaussian variable $N(0, 1)$. Let $\nu_0 = \mathrm{E}[-\log(1 - \Phi(X))]$, where $X \sim N(0, 1)$, *i.e.,* the expected score of a node from $V/S$. Similarly, let $\nu_1 = \mathrm{E}[-\log(1 - \Phi(X))]$, where $X \sim N(\mu, 1)$. Furthermore, let $\eta \triangleq \nu_1/\nu_0$. Since $\mu$ is a constant, we have $\nu_0$, $\nu_1$, and $\nu_1 - \nu_0$ are all constants. Also, $\eta$ is a function that grows with $\mu$. Our way of setting the function $c(\cdot)$ is straightforward: we set $c(e) = \omega \triangleq \frac{\nu_1 - \nu_0}{2}$ for all $e$.

We first show that when $S$ is substituted into (3), the objective value is at least $(1-\frac{\epsilon}{10})k\nu_1$. This can be seen by using a Chernoff type bound for the independent variables $-\log(1-\Phi(X))$ with $X \sim N(\mu, 0)$ (See Theorem I.1 in Appendix I for the statements and the proofs):

$$\Pr\left[\sum_{v\in S} b(v_i) \leq (1-\frac{\epsilon}{10})k\nu_1\right] \leq \exp(-\Theta(\epsilon^2\nu_1 k)). \tag{12}$$

Thus, the objective value is at least $E_s \triangleq (1-\frac{\epsilon}{10})k\nu_1 - \omega\cdot(k-1)$ with high probability. We next show that for any specific $S'$ of size $\ell$ such that $|S'-S|+|S-S'| \geq \epsilon k$, we have

$$\Pr\left[\sum_{v_i\in S'} b(v_i) - \omega(|S'|-1) \geq E_s\right] \leq \exp(-g(\mu)\epsilon^2\ell), \tag{13}$$

where $g(\mu)$ is a monotonic function in $\mu$. Then, using the fact that with probability $\geq 1-\epsilon$ the number of connected subgraphs of size $\ell$ is $\leq \frac{n}{\epsilon}(\tau_0)^\ell$ for some constant $\tau_0$, we can conclude that whp any $S'$ of size $\ell$ such that $|S'-S|+|S-S'| \geq \epsilon k$ cannot be an optimal solution.

We now move to prove (13). Our goal thus is to give a bound on the event $\sum_{v_i\in S'} b(v_i) - \omega(|S'|-1) \geq E_s$. The probability is maximized when $S \subset S'$. Let us write $S_2' = S' - S$. Thus, we shall find a bound for $\Pr[\sum_{v_i\in S} b(v_i) + \sum_{v_i\in S_2'} b(v_i) - \omega(\ell-1) \geq E_s]$.

When $\sum_{v_i\in S} b(v_i) + \sum_{v_i\in S_2'} b(v_i) \geq E_s + \omega(\ell-1)$, there exists a $\Delta \in \{-n, ..., n\}$ such that

$$\sum_{v_i\in S} b(v_i) \geq k\nu_1 + \Delta$$

and

$$\sum_{v_i\in S_2'} b(v_i) \geq -\frac{\epsilon\nu_1 k}{10} - \Delta + (\ell-k)\omega - 1.$$

Let us write $\Delta = \delta k\nu_1$ and $\ell = hk$. When $\Delta \geq 0$, we have

$$\Pr\left[\sum_{v_i\in S} b(v_i) = k\tau_1 + \Delta\right] \leq \exp(-c\delta^2\eta k\nu_0)$$

and when $\Delta \leq \Delta_m \triangleq \frac{\eta-1}{2}(\ell-k)\nu_0$:

$$\begin{aligned}
&\Pr\left[\sum_{v_i\in S_2'} b(v_i) \geq \frac{-\epsilon\nu_1 k}{10} - \Delta + (\ell-k)\omega - 1\right] \\
\leq\ & \exp\left(-\frac{c\left(\frac{\eta-1}{2}(\ell-k)\nu_0 - \frac{\epsilon}{10}\nu_1 k - \Delta\right)^2}{\ell\tau_0}\right) \\
\leq\ & \exp\left(-\frac{ck\nu_0}{h}\left(\frac{\eta-1}{2}(h-1) - \frac{\epsilon\eta}{10} - \delta\eta\right)^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Pr[\sum_{v_i\in S} b(v_i) + \sum_{v_i\in S_2'} b(v_i) - \omega(\ell-1) \geq E_s] \\
\leq\ & \sum_{\Delta\leq 0} \exp\left(-\frac{ck\nu_0}{h}\left(\frac{\eta-1}{2}(h-1) - \frac{\epsilon\eta}{10} - \delta\eta\right)^2\right) \\
&+ \sum_{1\leq\Delta\leq\Delta_m} \exp(-c\delta^2\eta k\nu_0) \\
&\quad\cdot \exp\left(-\frac{ck\nu_0}{h}\left(\frac{\eta-1}{2}(h-1) - \frac{\epsilon\eta}{10} - \delta\eta\right)^2\right) \\
&+ \sum_{\Delta>\Delta_m} \exp(-c\delta^2\eta k\nu_0).
\end{aligned}$$

It is not difficult to see that both the first and third summations $\leq \exp(-\Theta(\epsilon^2(\eta-1)^2\nu_0\ell/\eta))$. Therefore, it remains to analyze the second summation in the above inequality. Specifically, we want to understand when

$$\exp\left(-c\delta^2\eta k\nu_0 - \frac{ck\nu_0}{h}\left(\frac{\eta-1}{2}(h-1) - \frac{\epsilon\eta}{10} - \delta\eta\right)^2\right) \tag{14}$$

is maximized. One can see that the exponent is a quadratic function in $\delta$, which is maximized when

$$\delta = \frac{2(h-1)(\frac{2\eta}{5} - \frac{1}{2})}{\frac{\eta}{h}+1}.$$

When we plug in the optimal value of $\delta$ to (14), we have

$$\begin{aligned}
&\exp\left(-c\delta^2\eta k\nu_0 - \tfrac{ck\nu_0}{h}\left(\tfrac{\eta-1}{2}(h-1) - \tfrac{\epsilon\eta}{10} - \delta\eta\right)^2\right) \\
&\leq \exp(-\Theta(\epsilon^2(\eta-1)^2\nu_0\ell/\eta)).
\end{aligned}$$

Thus, (13) indeed holds.

## H Proof of Theorem 5.3

*Proof.* Let $\phi(x)$ be the pdf of $N(0,1)$. First, observe that for any non-negative functions $f$ and $g$ such that $\phi = f + g$, we may interpret a sample from $N(0,1)$ as a sample from the mixture of two distributions from $f$ and $g$ by using the following procedure:

- First let $F = \int_{-\infty}^{\infty} f(x)dx$ and $G = \int_{-\infty}^{\infty} g(x)dx$.
- Then with probability $\frac{F}{F+G}$, we draw a sample from the ditribution with pdf $\frac{f(x)}{F}$ and with probability $\frac{G}{F+G}$ we draw a sample from the distribution with pdf $\frac{g(x)}{G}$.

Let $\phi_\mu(x)$ be the pdf for $N(\mu, 0)$ and let $\tau = \frac{\phi(\mu)}{\phi_\mu(\mu)}$. We shall decompose $\phi(x)$ into a mixture of $\tau\cdot\phi_\mu(x)$ and $R(x) \triangleq \phi(x) - \tau\cdot\phi_\mu(x)$. We consider the following strictly simpler problem and give a lower bound on $\alpha$

for this problem: we still have the same setting that nodes from $V/S$ and from $S$ receive samples from different distributions and we are asked to find $S$. The only difference here is that when $v_i \in V/S$ receives a sample from $N(0,1)$, we assume the sample is generated from the mixture of $\tau \cdot \phi_\mu(x)$ and $R(x)$. Furthermore, when $v_i$ is sampled from $R(x)$, we also *explicitly label* $v_i$ as from $R(x)$. In other words, the algorithm knows the set of nodes that are sampled from $R(x)$. Notice that the new problem gives a strict superset of information and thus is information theoretically easier than the original problem.

Next, let us move to find a lower bound on $\alpha$ for the new problem. When a node is labeled as from $R(x)$, it is clear that the node should not be part of the output. It remains for us to find $S$ from the rest of the non-labeled node. But notice that all the rest of the signals are sampled from $N(\mu, 0)$. Thus, we cannot do better than randomly guessing. Since $\mu = \Theta(1)$, we have $\tau = \Theta(1)$. Thus, the size of the remaining unlabeled nodes is still $\Theta(n)$ (with high probability). One can see that in order to cover $\rho$ portion of nodes from $S$, the size of the final output has to be $\Theta(\rho n)$. □

# I Concentration bounds

In this section, we prove the following large deviations bound.

**Theorem I.1.** *Let $Y_i \sim N(0,1)$ and $X_i \sim N(\mu, 1)$. Let $\Phi(\cdot)$ be the cdf for $N(0,1)$. Let $\mu_x = \mathrm{E}[-\log(1-\Phi(X_i))]$ and $\mu_y = \mathrm{E}[-\log(1-\Phi(Y_i))]$. We have*

$$\begin{aligned}&\Pr\left[\left|\textstyle\sum_{1\le i\le n} -\log(1-\Phi(X_i)) - n\mu_x\right| \ge \epsilon n\mu_x\right]\\ &\le \exp(-c\epsilon^2 n\mu_x)\end{aligned} \tag{15}$$

*and*

$$\begin{aligned}&\Pr\left[\left|\textstyle\sum_{1\le i\le n} -\log(1-\Phi(Y_i)) - n\mu_y\right| \ge \epsilon n\mu_y\right]\\ &\le \exp(-c\epsilon^2 n\mu_y)\end{aligned} \tag{16}$$

*for a suitable constant $c > 0$.*

*Proof.* We shall prove the lower tail of (17), *i.e.,*

$$\begin{aligned}&\Pr\left[\textstyle\sum_{1\le i\le n} -\log(1-\Phi(X_i)) \le (1-\epsilon)n\mu_x\right]\\ &\le \exp(-c\epsilon^2 n\mu_x)\end{aligned} \tag{17}$$

for some constant $c > 0$. The other cases can be analyzed in a similar manner. Consider the moment generating function (mgf) of $-\log(1-\Phi(X_i))$, where $X_i \sim N(\mu, 1)$ and $\Phi(\cdot)$ is the cdf of $N(0,1)$. For convenience, we also denote $\bar\Phi(\cdot) = 1 - \Phi(\cdot)$ as the tail distribution of $N(0,1)$. We use the bound (Williams, 1991)

$$\bar\Phi(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \ge \frac{1}{\sqrt{2\pi}(x+1/x)} e^{-x^2/2} \tag{18}$$

for $x > 0$. We shall prove that the mgf of $-\log(1-\Phi(X_i))$ exists and is finite in a neighborhood of zero. Namely, consider

$$\begin{aligned}\phi(\theta) &:= E[e^{-\theta \log \bar\Phi(X_i)}]\\ &= E\left[\tfrac{1}{(\bar\Phi(X_i))^\theta}\right]\\ &= \textstyle\int_{-\infty}^{\infty} \frac{1}{(\bar\Phi(x))^\theta} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx\end{aligned} \tag{19}$$

Using (18), and by considering the region $\{x \le \eta\}$ and $\{x > \eta\}$ for some $\eta > 0$, the quantity (19) is bounded from above by

$$\begin{aligned}&\max\{\bar\Phi(\eta)^{-\theta}, 1\} \int_{-\infty}^{\eta} e^{-(x-\mu)^2/2} dx\\ &+ \int_\eta^\infty \left(\sqrt{2\pi}\left(x + \frac{1}{x}\right) e^{x^2/2}\right)^\theta \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx\\ =\ &\max\{\bar\Phi(\eta)^{-\theta}, 1\} \int_{-\infty}^{\eta} e^{-(x-\mu)^2/2} dx + (2\pi)^{(\theta-1)/2}\\ &\int_\eta^\infty \left(x + \frac{1}{x}\right)^\theta e^{\theta x^2/2 - (x-\mu)^2/2} dx\end{aligned} \tag{20}$$

Consider the second term in (20). For $0 < \theta < 1$, it is bounded by

$$(2\pi)^{(\theta-1)/2} \int_{x>\eta} C_1 x^\theta e^{\theta x^2/2 - (x-\mu)^2/2} dx < \infty \tag{21}$$

for some $C_1 > 0$, and for $-1 < \theta < 0$, it is bounded by

$$(2\pi)^{(\theta-1)/2} \int_{x>\eta} C_2 x^{-\theta} e^{\theta x^2/2 - (x-\mu)^2/2} dx < \infty \tag{22}$$

for some $C_2 > 0$. Therefore $(-1, 1)$ is contained in the domain of convergence of $\phi(\theta)$. This implies that $\phi(\theta)$ is infinitely differentiable in $(-1, 1)$. Define $\psi(\theta) = \log \phi(\theta)$ as the logarithmic mgf of $-\log(1-\Phi(X_i))$. The same convergence and differentiability behavior then holds for $\psi(\cdot)$ in the same region $(-1, 1)$.

To proceed, we use the Chernoff inequality

$$\Pr\left[\sum_{1\le i\le n} -\log \bar\Phi(X_i) \le (1-\epsilon)n\mu_x\right] \le e^{\theta(1-\epsilon)n\mu_x + n\psi(-\theta)} \tag{23}$$

for $0 \le \theta < 1$. Using the Taylor expansion $\psi(-\theta) = -\psi'(0)\theta^2 + \psi''(\zeta)\theta^2/2$ for some $\zeta \in (-\theta, 0)$, and the fact that $\psi'(0) = \mu_x$, (23) becomes

$$\begin{aligned}&\exp\left\{\theta(1-\epsilon)n\mu_x - n\psi'(0)\theta^2 + n\psi''(\zeta)\tfrac{\theta^2}{2}\right\}\\ &= \exp\left\{-\epsilon\theta n\mu_x + n\psi''(\zeta)\tfrac{\theta^2}{2}\right\}\\ &\le \exp\left\{-\epsilon\theta n\mu_x + n \sup_{u\in[-\theta,0]} \psi''(u)\tfrac{\theta^2}{2}\right\}\end{aligned}$$

For $\epsilon < N_1$, we choose $\theta = c_1\epsilon$, and the value of $0 < c_1 < N_2$ will be chosen small enough to get our result. Note that for $\epsilon < N_1$ and $c_1 < N_2$, any choice of $\theta = c_1\epsilon$ implies that $\sup_{u\in[-c_1\epsilon,0]} \psi''(u) \leq M$ for some constant $M > 0$. Hence (24) becomes

$$\begin{aligned} & \exp\left\{-c_1\epsilon^2 n\left(\mu_x - \sup_{u\in[-c\epsilon,0]} \psi''(u)\frac{c_1}{2}\right)\right\} \\ \leq \quad & \exp\left\{-c_1\epsilon^2 n\left(\mu_x - M\frac{c_1}{2}\right)\right\} \end{aligned}$$

Choosing $c_1$ small enough will then give $\mu_x - M\frac{c_1}{2} > 0$, which concludes the theorem. □

# J Missing calculations

## J.1 Proof of Equation 2

$$\begin{aligned} & \gamma^{k_0+\Delta k}\frac{1.55k(\tau_0)^\ell}{p\epsilon} \\ = \quad & \frac{1.55k}{p\epsilon}\gamma^{k_0+\Delta k}\tau_0^{k+\Delta k} \\ = \quad & \frac{1.55k}{p\epsilon}\gamma^{k_0+\Delta k}(\tau_0)^{k+\Delta k} \\ = \quad & \frac{1.55k}{p\epsilon}(\tau_0)^k\gamma^{k_0}(\gamma\tau_0)^{\Delta k} \\ \leq \quad & \frac{1.55k}{p\epsilon}(\tau_0)^k(\gamma^{\lambda\gamma})^k(\gamma\tau_0)^{\Delta k} \quad \text{(Using } k_0 \geq \lambda\gamma k\text{)} \\ \leq \quad & (2\tau_0\gamma^{\lambda\gamma})^k(\gamma\tau_0)^{\Delta k} \\ & \qquad \text{(using } \tfrac{1.55k}{p\epsilon} \leq 2^k \text{ for sufficiently large } k.\text{)} \\ \leq \quad & c_0^{-k} \end{aligned}$$

The last inequality holds when $\gamma^{\lambda\gamma} \leq \frac{1}{2c_0\tau_0}$ for any constant $c_0$.